Discussion

IPS 109: New Data Sources Meet HH surveys

Gero Carletto
Manager, Development Data Group, World Bank
Co-Chair, Inter-Secretariat Working Group on Household Surveys
As Siu-Ming clearly illustrates ...

• Big Data are attractive and promising for many reasons:
  ✓ Provide information on new phenomena (e.g. Type 1 and 2 scenarios)
  ✓ Improve the accuracy of estimates coming from traditional surveys (typically by increasing estimators efficiency, e.g. Type 3 and 4 scenarios)
  ✓ Save data collection costs, reduce respondents burden, increase timeliness, improve temporal and spatial resolution of statistical information, etc.

• However, Big Data cannot be used “as is”!
  ▶ Big Data selectivity would typically lead to biased estimation

• Suitable methods must be developed to address Big Data selectivity (and many other possible Big Data issues):
  ▶ Data Integration stands out as one of the most promising approaches
DI approach to Big Data

• DI is an umbrella term: methods differ in underlying assumptions and/or technical implementation details, but share a common idea:
  
  ▶ Integrate Big Data with information (micro or macro) coming from datasets whose representativeness is “guaranteed” (e.g. probability samples, census data, statistical registers)
  
  ▶ To that objective, exploit auxiliary “bridge” variables $X$ that are able to:
    a) explain the mechanism of selectivity of Big Data
    b) predict well the variables of interest $Y$

• The integration process should allow Big Data to “borrow representativeness” from traditional datasets, resulting in the sought-after bias reduction effect
DI: 4 Data structure Scenarios

• DI methods illustrated for **Type 1 and 2 scenarios** (i.e. no Y variable available in the traditional dataset A) based on:
  
  ► **Attach suitable weights** to observations in the Big Data sample B, then proceed with weighted estimation *using B*
    – Techniques: *propensity modeling* and inverse propensity weighting, *(pseudo-)calibration*, etc.

  ► **Transport Y values** from the Big Data sample B to the traditional dataset A, then *use* the completed (synthetic) *dataset A* for estimation
    – Techniques: *massive imputation* (i.e. model and fit $Y \sim f(X, \Theta)$ on B, then use the fitted model to predict and impute $Y^*$ values in A), *sample matching* (i.e. impute Y values in A from nearest-neighbor donors in B, using X as matching variables), etc.

• The **Doubly Robust Estimator** smartly blends the above approaches to gain protection against bias when the *modeling assumptions* underpinning either of the two fail **(but not both!)**
DI: 4 Data structure Scenarios

• To no surprise, DI methods illustrated for **Type 1 and 2 scenarios** are akin to approaches used - before the advent of Big Data - to address selectivity issues in “traditional” survey workflows, e.g. treatment of non-response in probability surveys.

• For **Type 3 and 4 scenarios** (i.e. when the Y variable is available in both the Big Data sample B and the traditional dataset A), the goal of DI methods is to:
  ▶ Increase the efficiency of estimates derived from the traditional dataset A, by using suitable estimates derived from Big Data B as benchmarks.

• Of course, the goal above must be achieved **without importing into A the selection bias affecting B**, which makes the problem highly non-trivial.
  ▶ This is exactly the purpose of the **Regression Data Integrator (RDI)** approach proposed in Kim and Tam (2020)!
The RDI approach

• Note that in the RDI approach:
  ▶ The “traditional” representative dataset A still plays a crucial role in adjusting for selection bias of the Big Data sample B
  ▶ In return, the Big Data sample B provides for variance reduction

• The ability to express the RDI as a calibration estimator allows Kim, and Tam (2020) to extend the method to cover more realistic and complex applications, e.g.
  ✓ Inclusion of covariates (observed in both A and B, or only in B)
  ✓ Duplicated units in B
  ✓ Measurement errors affecting variable Y in A and/or B
  ✓ Non-response in A
  ✓ Linkage errors in identifying units belonging to A∩B

• Given their potential impact, all the above theoretical extensions of the RDI need and deserve empirical evidence from real-world applications and/or from simulations
Thoughts for low- and middle-income countries ...

- Besides methodological breadth and depth, Siu-Ming’s presentation makes us even more aware
  ✓ of the **statistical challenges** that come along with Big Data promises
  ✓ of the **key and enduring role of traditional surveys** in addressing these challenges

- Both the points above have serious implications for low- and middle-income income countries:
  - As these countries typically suffer from chronic “*traditional data deprivation*”, they would seem best positioned to receive greatest benefits from Big Data
  - However, deriving sound inferences from Big Data requires to leverage auxiliary variables that must come from high-quality traditional data (often not available)
  - It seems that these countries risk (once again) to be trapped in a **vicious cycle, leading to greater inequities**
Thoughts for low- and middle-income countries ...

• While necessary, rich and high-quality traditional data are not sufficient to take advantage of Big Data. Other enabling factors are needed.

• Proper treatment of Big Data calls for a **wide range of skills and capabilities**
  
  i. **Knowledge of domains and phenomena Big Data refer to**
  
  ii. **Understanding of the data generating mechanism of Big Data sources**
  
  iii. **Statistical modeling abilities, to exploit i. and ii. for inference**
  
  iv. **Survey design, planning and management** (e.g. sometimes a dedicated ancillary sample survey may be required to observe units not covered by the Big Data source and allow propensity modeling)
  
  v. **Data science capabilities, to prepare Big Data for analysis** (e.g. to extract/build the variables X needed for data integration, which typically do not exist as such in unstructured Big Data)
  
  vi. **Mastery of techniques and tools of the Official Statistics field** (e.g. for calibration, imputation, statistical matching, record linkage, ...)

---

8
Thoughts for low- and middle-income countries ...

- Going beyond statistical methodology, **further requirements** come to mind as enabling factors to integrate Big Data in statistical production:
  
a. **Dedicated ad strengthened IT infrastructures (both HW and SW)**, for Big Data access, acquisition, storage, processing and analysis

b. **Multidisciplinary teams** (successful Big Data project always involve domain experts, statisticians, data scientists, ICT experts, and more)

c. **Partnership with Big Data providers** (e.g. WB Development Data Partnership [http://datapartnership.org/](http://datapartnership.org/))

d. **Institutional commitment of the NSO** (Big Data projects can face higher levels of skepticism and risk of failure than ordinary projects)

e. **Partnerships across NSOs** (eg ESCAP CoP)

f. **Legal framework for data access and data privacy**
In conclusion ...

• The international community should put substantial effort to “leveling the playing field” for low and middle-income countries to prevent Big Data from becoming yet another form of divide

• Since low- and middle-income countries are still lagging behind in many of the “pre-conditions” to fully exploit the opportunities offered by alternative data sources, dedicated capacity building and infrastructure investments are paramount to ensure an equitable data transformation

• What is needed is what the WB World Development Report *Data for Better Lives* calls a “social contract for data”